OFFICE OF ELEMENTARY AND SECONDARY EDUCATION

SCHOOL SUPPORT AND ACCOUNTABILITY

2023 State Assessment Conference
1C - Fairness & Comparability

September 26, 2023

FOCUS AREA: ASSESSMENT PEER REVIEW

1C - Fairness & Comparability

Phoebe Winter, Independent Consultant, phoebe.winter@outlook.com
Brooke Nash, ATLAS, bnash@ku.edu
Zach Warner, NYS Ed Dept, zachary.warner@nysed.gov
Meagan Karvonen, ATLAS, Karvonen@ku.edu

A Note About this Conference/Session

- The purpose of this conference/session is to provide an opportunity for State education agency (SEA) staff to interact and engage with relevant experts and other SEA staff about the Department's assessment peer review.
- The observations and opinions of the session presenters are their own.

Session Overview

- Overview
- State and Consortium Examples
- Responding to Peer Review Requirements
- Q&A

Overview

Fairness and Comparability: Considerations for Multiple Approaches

Newly adopted approaches to assessment typically differ from traditional approaches in the inferences are designed to support. They may introduce between-student variation in:

- What is assessed.
- When students are assessed.
- How they are assessed.

Examples of different approaches:

- Matrix sampling
- Through-year and instructionally embedded assessments
- Performance assessments

Multiple Approaches Handout

Download for reference:

- Summary of assessment programs featured in Focus 1 sessions
- Focus 1 slides (excerpt from plenary session)

Critical Elements (CEs) Addressed

4.2 Fairness and Accessibility.

4.5 Multiple Assessment Forms.

4.6 Multiple Versions of an Assessment.

Also applicable to test design and development (2.1), validity (3.1), score reporting (6.4)

CE 4.2 Fairness and Accessibility

For all State academic and ELP assessments, assessments should be developed, to the extent practicable, using the principles of universal design for learning (UDL) (see definition).

For academic content assessments, the State has taken reasonable and appropriate steps to ensure that its assessments are accessible to all students and fair across student groups in their design, development and analysis.

CE 4.5 Multiple Assessment Forms

If the State administers multiple forms of academic assessments within a content area and grade level, within or across school years, the State ensures that all forms adequately represent the State's academic content standards and yield consistent score interpretations such that the forms are comparable within and across school years.

CE 4.6 Multiple Versions of an Assessment

If the State administers any of its assessments in multiple versions within a subject area (e.g., online versus paper-based delivery; or a native language version of the academic content assessment), grade level, or school year, the State:

- Followed a design and development process to support comparable interpretations of results for students tested across the versions of the assessments;
- Documented adequate evidence of comparability of the meaning and interpretations of the assessment results.

Defining Fairness and Comparability

- Fairness
  - All students have the opportunity to demonstrate the targeted knowledge, skills, and understandings.
  - The assessment supports valid inferences that are comparable across the tested population.
- Comparability
  - Scores support inferences at the desired score level(s).
  - Scores support inferences at the desired aggregation level(s).
- For both
  - Content measures proficiency according to the same construct.
  - Results can be used for the same purposes regardless of test form or test conditions.

Fairness is a necessary condition for score comparability.

How should fairness and comparability be considered in newly adopted assessment approaches?

Design and Development

- Accessibility Built In
- Clear and Fair Soring Rules
- Multiple Forms Created to be Comparable

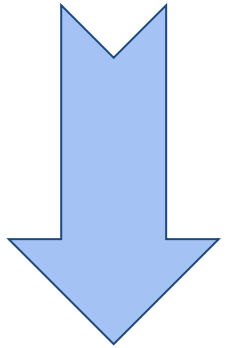
Fairness and Accessibility

- Opportunity to Learn (OTL)
  - Procedures
  - Depth and breadth of content taught matches depth and breadth of content assessed.
  - Timing of assessment relative to instruction is similar for all students.
- Accessibility Tools
- Accommodations

How do purpose and use affect how a state attends to comparability?

What is/are the target inference(s)?

How will results be used?



How do I define and design for score comparability?

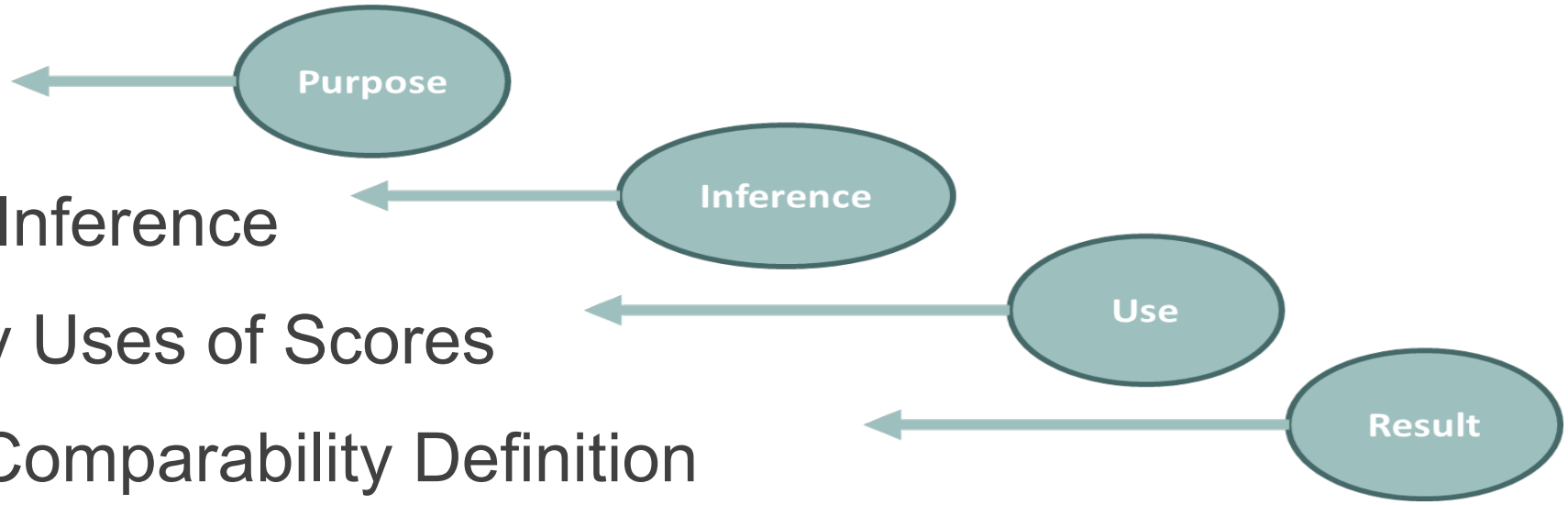
An Example

Goal

Target Inference

Primary Uses of Scores

Score Comparability Definition



Constructing Comparability

- Assessment design, e.g.,
  - Content sampling rules.
  - Administration mode(s).
  - Tools…
- What to standardize, e.g.,
  - Delivery
  - Familiarity with conditions.
  - Access to what is being asked.
  - Opportunity to respond…

In General, an Approach that Yields Comparable Results Will…

- support valid inferences for all students in the target population.
- have student forms designed to all be as closely aligned to grade level content standards as other forms.
- provide results that are equally as reliable at the score level at which inferences are made.
- classify students into achievement levels based on the same degree of knowledge and skills.

State and Consortium Examples

Examples

How should fairness and comparability be considered in newly adopted assessment approaches?

- Instructionally embedded assessments
- Performance assessment

Instructionally Embedded (IE) Assessment - (1)

- **Fairness in Design**
  - Features:
    - "Year-round" administrations that provide comparable contexts to demonstrate knowledge, skills and understandings (KSUs).
    - Flexibility in administration
      - when and what to assess.
      - ensuring test security.
    - Balancing standardization with accessibility.

Instructionally Embedded Assessment - (2)

- **Fairness in Design (continued)**
  - Example evidence
    - High-quality and comprehensive test administration training and manuals.
    - Documentation of allowable practices.
    - Test administration observations and student cognitive labs.

Instructionally Embedded Assessment - (3)

- **Fairness in Development**
  - Feature
    - Large item pools to support embedded approach.
    - Use of ECD-based task templates to ensure items are fair across student groups.
  - Example evidence
    - Descriptions of ECD-based task template models and their development process.
    - Descriptions of and sample materials from item writing training, including how fairness was addressed.

Instructionally Embedded Assessment - (4)

- **Forms Represent Content Standards**
  - Feature
    - Short embedded assessments are dynamically generated and may only assess 1-3 standards at a time.
  - Example evidence
    - Description of assessment assignment procedures to ensure intended coverage and methods for monitoring assignment procedures and coverage during window.
    - Analysis of item pool to show content available across breadth of standards.

IE Assessment - (5)

- **Forms Yield Consistent Score Interpretations**
  - Feature
    - Use of evidence centered design (ECD) based task templates to ensure items are written to precise cognitive specifications
    - Scoring model and methods consistent with test design and intended use of results
      - For example, a diagnostic classification model that provides student mastery statuses of each assessed skill.
  - Example evidence
    - Item data review including evidence that items written to the same knowledge, skills, and understandings perform similarly.
    - Evidence of model fit (item fungibility assumption).

Performance Assessment

- NY has a longstanding practice of including performance-based items in science examinations at all levels.
- The science performance items allow students to demonstrate specific knowledge and skills articulated in the learning standards through hands-on laboratory experiences. These items make up 15% of the total test score, combined with a written test.
- Although not currently planned to be used for accountability purposes under ESSA, NY's Performance-Based Learning and Assessment Networks (PLAN) Program is exploring the potential for New York's educational assessment strategy to be reimagined in a way that purposefully fosters high-quality instructional opportunities, provides authentic measures of deeper learning, and better prepares students for college and the workplace.

Performance Assessment

Fairness & Comparability Considerations in Design

- Focus on target(s) of measurement, rather than task type.
- Universal Design for Learning (UDL) considerations should start right at the beginning, not wait until task development (if using multiple versions, consider if there are limitations in how content can be presented).
- Balance flexibility in assessment with need for consistent interpretations as well as access for all students.
- Consider limitations of assessment users too (e.g., what materials are needed/available locally.

Performance Assessment

Fairness & Comparability Considerations in Development

- UDL should drive task development (which will flow naturally since it was considered in the design phase).
- Focus is on the knowledge/skills to be measured (i.e., for comparability), minimizing rigid requirements where possible.
- Documentation/specification!
- Major key is ensuring that students have what they need to show what they know and are able to do; not to ensure that everyone gets the same presentation and options.

Performance Assessment

Fairness & Comparability Considerations in Scoring/Score Use

- Scoring should focus on targets of measurement and, as much as possible, ensure that other aspects do not influence results (e.g., if language choice is not a key target, it should not impact scores).
- As with all assessments, score use should be tied to purpose.
- If interpretations are impacted by the flexibility in the assessment, this should be clear in reporting.
- Depending on purpose/use, comparability may not be a reasonable goal; instead, inferences would be standards-based/criterion-referenced.

Responding to Peer Review Requirements

Preparing for Peer Review

- Purpose and use
  - Clearly describe purpose and use.
  - Explicitly tie evidence for relevant CEs to purpose and use.
- Using scores
  - Define how scores are used to meet Title I requirements.
  - Tie level of comparability to how scores are used for Title I.
- Uses outside of peer review purview
  - If any non-Title I uses affect how a CE is addressed, explain as needed.
  - Clearly delineate non-Title I uses if they are discussed.

Goal: Ensuring Peer reviewers understand the system and reasons for using evidence that may be atypical.

Evidence of Fairness and Comparability

- Provision of appropriate supports and tools to students taking assessments
- Appropriate Accommodations (empirical studies, literature reviews)
- Professional judgment in design/development phase
- Technical properties of the assessment under different testing modalities
- Audits
- Score comparisons across forms
- Relationships to other measures
- Cognitive processes

What are some misperceptions of peer review fairness and comparability requirements that you have encountered?

Discussion Questions

- What are the biggest challenges in providing evidence of fairness and comparability?
- How do these challenges affect how you might address 4.2 and 4.5 for new/different assessment approaches?
- What clarifications would a state need to what's in the guidance on 4.2 and 4.5 to address these challenges?

QUESTIONS?

Still more questions?

- Submit your questions using the QR code
- Attend session 1G (*Preparing for Assessment Peer Review*) Wednesday afternoon for answers